Understanding Usage and Activity in Cellular Networks by Investigating HTTP Requests

Fehmi Ben Abdesslem
SICS Swedish ICT
fehmi@sics.se

Anders Lindgren
SICS Swedish ICT
andersl@sics.se

Andrea Hess
SICS Swedish ICT
andreah@sics.se

Abstract— **The number of mobile devices is estimated to now exceed the world's population, using more and more cloud services, and hence generating more and more traffic. Smartphones generate 95% of the total global handset traffic, and while approximately half of this traffic is sent to cellular networks, other handsets such as tablets are also using increasingly the cellular networks. This paper provides a closer look at the traffic generated on cellular networks by exploring billions of HTTP requests sent by millions of users to a nation-wide cellular network during 41 days. We confirm that – as in many other contexts – 20% of the users are responsible for more than 80% of the requests and provide a deeper analysis of the cellular network usage. Furthermore, we characterise the activity of users on their mobile device and which cloud services they use. For instance, almost 30% of the users use the cellular network frequently, mainly using search services and social networks, but 20% of their requests are sent to advertisement and tracking systems.**

I. Introduction

Over the past decade, we have seen a rapid transformation in the way we access information and network services. With the advent of smart phones and cheap cellular data plans, Internet access has gone from being an office or home utility to being a ubiquitous part of the way we live our daily life. Additionally, cloud revolution has changed the way we use our devices and use applications and services by moving more and more from the local device to data centers in the cloud. This trend has recently also moved into the cellular domain and users access many cloud services from their mobile devices. With these trends, there has been an explosive growth in the amount of data sent over cellular networks. Global mobile data traffic grew 81% in 2013, and more than half of this traffic is transported by cellular networks. With the number of mobile devices now exceeding the world's population, mobile data traffic is expected to further grow 11-fold by 2018 [1].

This is becoming an increasing problem for cellular network operators and cloud service providers as the extensive mobile data traffic causes capacity problems: in terms of traffic volume, but also in terms of application signalling that keeps the network active and prevents power-saving features. Understanding the usage patterns on mobile devices is thus becoming more and more important, in order to better adapt the network and cloud infrastructure and mechanisms to users' needs and behaviours.

In this paper, we use a large dataset containing HTTP requests collected by a major cellular operator from most of their mobile customers over the country for 41 days. The dataset contains tens of billions of requests generated by millions of users. To the best of our knowledge, this is the first study to look at such a large dataset to characterise user behaviour in cellular networks. We consider three main aspects of analysis. First, the usage patterns of users in the cellular network is characterised. Users are classified based on their frequency of generating cellular data traffic indicating that many users only access the Internet on their mobile device occasionally, but that the majority of all traffic comes from frequent users who use their mobile device almost every day. Furthermore, temporal usage patterns and types of devices used are investigated. Second, a more in-depth analysis of the type of user activity in the network is conducted. The most popular cloud services accessed and applications used are identified and assigned to categories in order to get an overview of the activities a user is engaged in. Finally, the correlation between temporal usage patterns and activity types are studied to understand how the usage of different services and applications vary by the time of day.

The remainder of this paper is organised as follows. Before describing the dataset in Section III, an overview of related works is presented in Section II. The dataset is then analysed in two parts. We first show how users generate traffic in Section IV, by characterising the usage frequency and patterns. Then Section V describes what they generate, and provides an insight on their activity. Section VI discusses the characteristics observed in the dataset and concludes this paper.

II. Related Work

While mobile calling patterns have been extensively studied by analysing CDR (Call Detail Records) data (see [2], [3]), understanding IP-based cellular network traffic is still an open challenge. Studying such traffic helps optimising content caching or capacity planning, for example. In this section, we present related works in this field in two parts: traffic analysis, and network user profiling.

Related work targeting traffic analysis can be found in [4]–[7]. In [4], correlations between different applications and basic network metrics (volume, inter-arrival time, etc.) on packet, flow, and session level are discussed for one-week GPRS trace data, and compared to the traffic caused by the same applications in a wired network. In [5], a prediction model is proposed to infer traffic volume from device types (wireless modem, smart phone, etc.) and associated application usage patterns (for categories, e.g., mail, browser). In [6], the authors characterise popularity, geographic coverage, and usage periodicity of apps based on one-week trace data of a UMTS network. The finding that 20% of popular applications have only local coverage emphasises the potential of content caching for network service providers. Finally, the authors in [7] study temporal and spatial traffic variations based on one week of nation-wide cellular trace data. They observe periodic patterns in terms of generated traffic volume, activeness, and re-appearance at locations. However, this periodicity is less recognisable when looking at individual stations than when looking the overall network.

Related work targeting network user profiling are found in [8]–[12]. In [8], hierarchical co-clusters of users and browsing profiles are built for relatively short time intervals (between 0.5 h and 6 h) from one-day network traces. The browsing behaviour of 500K users is captured and more than 60% of the users do not change their behaviour during the observation period. In [9], authors derive correlations between people's application interests and location visiting patterns of 280K users from a 3G cellular dataset, during one week in a metropolitan area. A key observation is that the most popular applications in the users' "comfort zones" (comprising their mostly visited locations) are of the music category, while users leaving this zone tend to use less battery intensive apps and to stay connected via social and communication apps. Studies of internet usage patterns for small groups of (up to 255) users based on very detailed data captured directly on the mobile phones have been presented, e.g., in [10], [11]. The authors in [12] analyse phone application usage with high granularity over four months for a larger set of users. The data have been probed by means of an application installed by more than $4,000$ users located mostly in the US and Europe, and thus, overcome bias problems arising when handing probing devices out to a small, homogenous user base (e.g., students of the same faculty). Despite their smaller scale, device-level probing studies are not limited to applications accessing the network and allow thus to investigate dependencies to offline applications.

Our work is complementary to these previous studies and provides further insights on the traffic patterns and user activities in a larger dataset.

III. Dataset

The analysis presented is based on a large-scale dataset of cellular network traffic traces from a major European operator. The traces were collected on a national scale over a period of 41 days, from 16th December 2011 at 17:00 to 25th January 2012 at 18:00. The dataset contains the URLs of several tens of billions HTTP requests together with timestamps, content sizes, and anonymised identifiers of the users and the cells.

Out of the 961 hours between the first and last request collected, 9 hours are missing from the dataset: two hours on 21st December, three hours on 2nd January, two hours on 22nd January, and one hour on 23rd and 25th January. Although those missing hours are noticeable, for instance, when looking at the amount of requests on the days with missing hours, we believe that they are not changing any of our conclusions.

IV. Usage Characterisation

In this section, we look at general aspects of network usage regarding frequency and time of usage as well as device types.

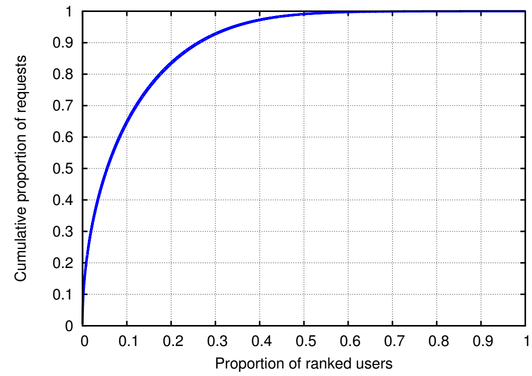

Fig. 1. Cumulative Distribution Function of the requests. 20% of users generate more than 80% of the requests.

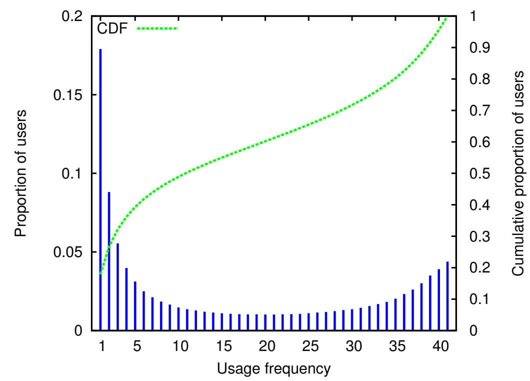

Fig. 2. Distribution of the users according to their usage frequency. Around 17.8% of the users appear only one day in our dataset, whereas approximately 4.4% appear every day.

A. Usage Frequency

The traffic pattern of mobile devices depends on the device type, the data plan subscribed by the user, the availability of WLAN networks that would be used to offload part of the traffic, and of course, the user interest for services requiring access to the network. All these factors explain the heterogeneity of behaviours observed in the dataset.

Figure 1 shows the distribution of the users depending on the number of requests generated. This typical distribution is found in several other contexts, and shows that the Pareto rule holds in the dataset: 20% of the users generate more than 80% of the requests.

We define the usage frequency $f(u)$ for a user u as:

$$f(u) = \sum_{d=1}^{41} D_u(d),$$

where $D_u(d) = \begin{cases} 1 & \text{if } u \text{ sends a requests in day } d \\ 0 & \text{otherwise} \end{cases}$

The usage frequency expresses the number of days in which users have sent at least one request. Since all users identified in the dataset have sent at least one request within the 41-day collection period, they are all active at least one day, hence we have $1 \leq f(u) \leq 41$. Figure 2 shows the distribution

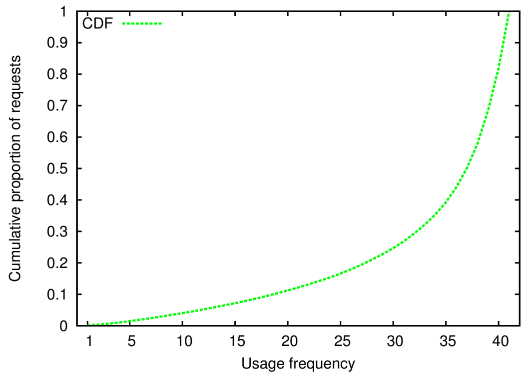

Fig. 3. Distribution of requests according to usage frequency of their users. Around half of requests are generated by users appearing at least 37 days.

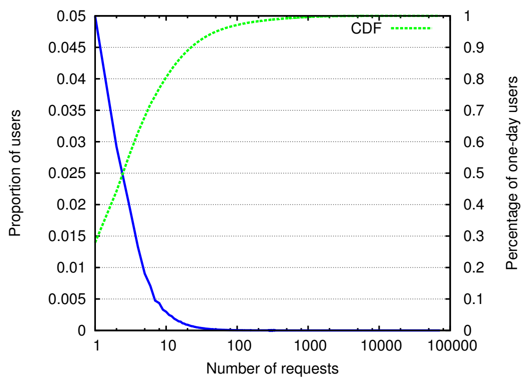

Fig. 4. Distribution of one-day users per number of requests. Around 80% of one-day users only send less than 10 requests on their active day.

of users according to their usage frequency. Out of the millions of unique users identified in the dataset, almost half have a low usage frequency of 10 days or less. We refer to them as *sparse users*. Among sparse users, around 17.8% of the total users appear only one day in the dataset. The number of users first decreases with the usage frequency, but then increases when the usage frequency is greater than 21 days. There is an important group of users, around 29.6%, with a usage frequency greater than 30 days. We refer to them as *frequent users*. Users that are neither sparse nor frequent users are referred hereafter as *moderate users*.

Sparse users clearly outnumber the other user categories, but because they are less active, they are individually expected to generate less traffic. On the other hand, as a large category, the cumulated traffic may potentially be more important than traffic generated by smaller categories.

Figure 3 depicts the distribution of the requests according to the usage frequency of the users. It clearly shows that the number of requests grows with the usage frequency. Despite outnumbering other categories, the sparse users only generate 4% of the requests, while moderate users, the smallest category, generate more than four times more. The users who appear only once in the dataset are generating 0.1% of the requests, and Figure 4 shows that most of them are sending less than 10 requests, and mainly to the ISP portal. This is probably explained by a user mistake or a user trial of the browsing feature, which might be pre-configured to display

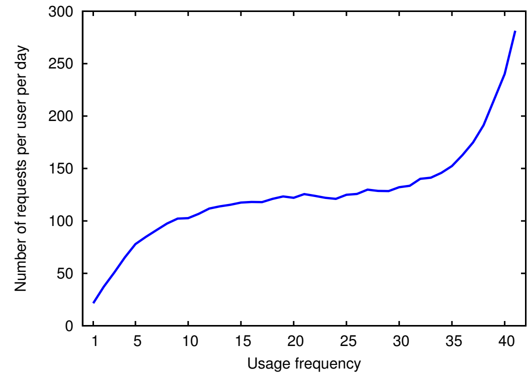

Fig. 5. Number of requests per user per day depending on the usage frequency $f(u)$. Users with a low usage frequency generate less requests during their active days than frequent users. The difference between moderate users themselves, ranging from 10 to 30 active days, is less important.

the default portal. Finally, frequent users are responsible for most of the traffic, with 77.2%.

We identified two users who generated an unexpectedly large number of requests, representing around 6% of all requests. This is due to collection problems and these requests were actually generated by different users. Since we cannot identify the real users sending these requests, we discard them in user-related analysis.

The number of both users and active days have an influence on the number of requests generated by each user group. By looking closer into users in different categories and on a single day, we are interested in understanding the difference, if any, between their request volume. Figure 5 shows the number of requests per user and per day for different usage frequencies. The graph shows that the higher usage frequency, the more requests users send on average on their active days. Hence, the higher volume of requests observed on frequent users is not only due to their higher number of active days, but they also generate more requests on each active day. Table I summarises the differences between the categories: sparse users form the largest category, but most of the requests are generated by frequent users, who do not only generate more requests per user, but also generate more requests per active day.

TABLE I. CHARACTERISATION OF USER CATEGORIES.

	Sparse users $f(u) \leq 10$	Moderate users $10 < f(u) < 30$	Frequent users $f(u) \geq 30$
% of users	48.9%	21.5%	29.6%
% of requests	4.0%	18.7%	77.2%
Req/user*	230	2,444	7,287
Req/user/day*	73	121	176

* average values

B. Time of Usage

To learn about the temporal variations in usage, we look at the usage intensity during each day. Figure 6 depicts the proportion of unique users and requests per hour during the course of a day. The peaks for both users and requests take place between 17:00 and 19:00, while in the early morning hours the values deviate considerably. For example, at 4:00 in the morning about 50% less users than at the peak hours

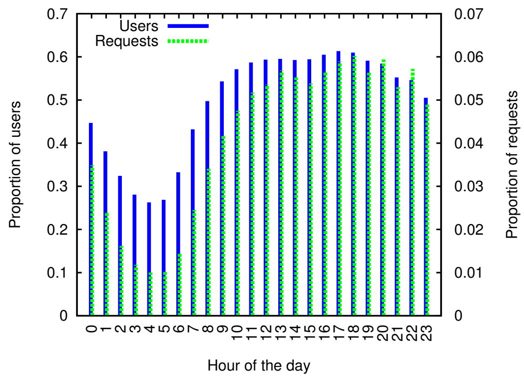

Fig. 6. Cumulated number of users and requests over the collection period, for each hour of the day.

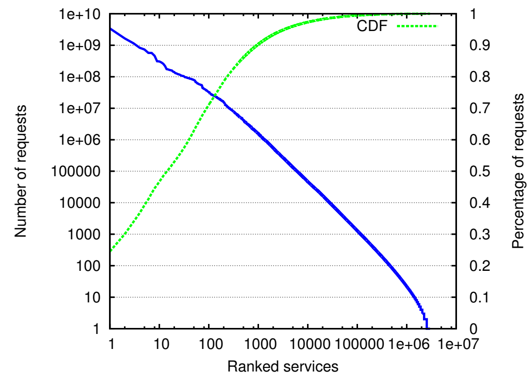

Fig. 7. Distribution of requests per service. The top 1,000 services are targeted by 90% of the requests, and the top 100 services by 71% of the requests.

send requests, whereas the number of requests amounts to only about 15% of the peak load. The relatively low number of requests might indicate automatic messages sent to a server in large time intervals or software updates done by an application automatically at this daytime.

C. Devices Used

The dataset does not indicate the type of device generating a request. However, some online services require the app or browser to send information about the device. We manually explored the content of the URLs generated by frequent users and identified services such as Google Search, Google Play, trackers and advertisement platforms. The operating system of more than half of the frequent users can be identified. Table II shows the number of devices identified as running different operating systems and the corresponding request count.

TABLE II. OPERATING SYSTEMS OF FREQUENT USERS.

Operating System	Number of devices	Number of requests
iOS	773,285	6,705,069,660
Android	604,571	4,690,797,911
Windows	108,037	1,557,879,060
Multi	33,972	500,107,436
Blackberry	15,369	76,520,756
Symbian	4,430	31,956,360
Unknown	1,426,110	8,050,759,555

V. ACTIVITY CHARACTERISATION

In this section, we categorise the services and apps utilised and investigate the extent and the temporal variations of their use.

A. Services

There are $11,282,008$ domains targeted by the requests. Domain names with a same second-level domain are considered as offering the same service. For instance, requests sent to `*.facebook.com` are all considered as related to the same service, facebook. $2,962,420$ services are targeted, either from a web browser, or from an application. However, Figure 7 shows that the distribution of the requests is heavily skewed: the top $1,000$ services (the top 0.03%) are targeted by 90% of the requests. Even the top 100 services are targeted by as much as 71% of the requests, which is more than 21 billion requests.

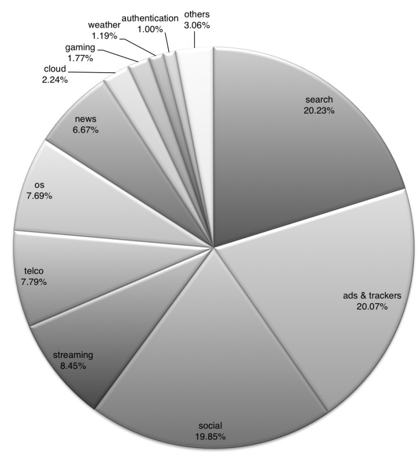

Fig. 8. Pie chart of requests sent for each popular service category by frequent users. Most of the requests are sent for search, ads & trackers, social, and media services.

Frequent users are mainly responsible for 80% of these 21 billion requests to the top 100 services, therefore the following analysis focuses on the traffic generated by frequent users. These services can be classified in 16 categories, described in Table III.

TABLE III. SERVICE CATEGORIES.

Category	Description	Example
search	Search engines and mapping services.	`bing.com`
ads & trackers	Advertisement systems and stat trackers.	`doubleclick.com`
telco	Service specific to the ISP.	`isp-portal.com`
social	Online social networks.	`facebook.com`
streaming	Audio and video streaming services.	`youtube.com`
cloud	Cloud services.	`amazonaws.com`
news	News websites and blogs.	`bbc.co.uk`
os	Services specific to the OS.	`windowsupdate.com`
authentication	Authentication and security services.	`verisign.com`
software	Software download websites.	`adobe.com`
gaming	Gaming websites or servers.	`gameloft.com`
weather	Weather forecast services.	`accuweather.com`
adult	Adult websites, pornography.	`youporn.com`
shopping	Online trading websites.	`ebay.com`
road service	Satnav and taxi services.	`tomtom.com`
dating	Online dating services.	`match.com`

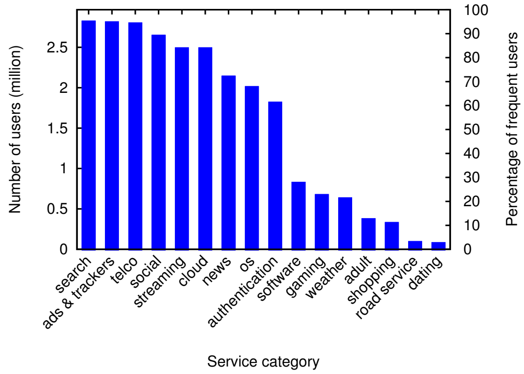

Fig. 9. Number of unique frequent users for each service category. Services in categories such as search, ads & trackers, or social are used by more than 80% of frequent users, whereas services such as weather, adult, or dating are only used by less than 25% of frequent users.

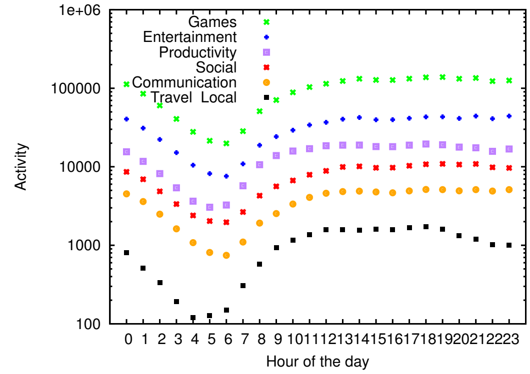

Fig. 10. Category of apps used for each hour of the day. One unit of activity represents one user using an app from a given category during that hour.

The amount of traffic generated by different service categories depends on (*i*) the nature of the service itself, as some services require accessing the servers more frequently, (*ii*) the usage frequency of the service, as some services are more frequently used than others, and (*iii*) the number of users of these services as a popularity indicator. Figure 8 shows the distribution of the requests sent by frequent users to the top 100 services, for each category. Search, ads & trackers, and social categories are generating the most traffic in the dataset. However, Figure 9 shows that categories such as cloud and streaming with a low number of requests, are used by a high number of users (more than 80% of frequent users).

B. Apps

The services described in the previous section are identified by looking at the HTTP requests that are generated by users accessing these services. This analysis allows us to identify online services that always require Internet access, and in particular sending HTTP requests to function. However, much user activity on smart mobile devices takes place on various downloaded applications (”apps”), such as popular games, health and fitness applications, and productivity tools. Unfortunately, by only looking at services that generate HTTP requests as part of their operation, we miss all these other types of activity. Analysis of HTTP requests to online application catalogs such as Google Play has shown that the types of apps downloaded show quite different characteristics than the types of services generating data traffic [13]. For example, the most popular category of downloaded applications is games, while Figure 8 shows that gaming only contribute to a very small fraction of data traffic generated.

It is however possible to identify some app usage even when applications do not utilise the Internet as part of their normal operation. This can be done by exploiting information sent to ads & tracker services for free versions of apps that use ads to generate revenue instead of charging for downloads. To track the display of ads (for billing purposes), the apps send requests to the ad tracking service. A closer look at the requests in this category provides an insight on the usage of these apps. In particular, there are 734 million requests to the advertisement platform *Doubleclick*[1] in the dataset, of which 382 million (52%) contain the name of the Android app generating it. Unfortunately, since URLs longer than 128 characters are truncated in our dataset, only 288 million of the requests contain the complete app name. In total, 79,576 unique apps were identified, which were then categorised, using the standard categories in Google Play as shown in Table IV. The table additionally shows the number of downloads from Google Play for apps in each category seen in the dataset.

TABLE IV. APP CATEGORIES.

Category	Number of apps in category	Number of Doubleclick users	Android Market downloads
Games	10,562	300,430	2,769,311
Removed	28,907	221,061	1,819,255
Tools	6,410	175,997	1,284,762
Entertainment	5,079	142,060	777,023
Music & Audio	1,940	101,067	640,835
Productivity	2,180	68,397	367,131
Lifestyle	2,250	45,595	204,808
Photography	998	32,734	163,386
Social	969	31,018	505,016
Sports	1,220	23,255	110,171
Communication	1,664	21,002	629,939
Personalization	4,218	20,327	382,006
Health & Fitness	1,135	19,133	109,846
News & Magazines	1,268	19,120	236,487
Media & Video	1,311	16,990	223,461
Weather	403	13,655	155,161
Transportation	594	11,201	205,537
Education	1,730	10,754	60,153
Travel & Local	2,078	10,663	1,085,340
Books & Reference	1,762	10,368	87,042
Business	780	9,583	37,549
Medical	388	7,995	19,202
Finance	717	7,075	114,850
Comics	258	3,934	13,681
Shopping	489	1,574	198,321
Libraries & Demo	266	558	32,013
Total	79,576		12,232,286

C. Temporal Activity Patterns

A problem with studying the apps that are extracted from the doubleclick requests in the dataset is that the frequency and intensity of requests from a particular app to the ad tracker does not necessarily correlate with the intensity of usage for that app. In order to further analyse the usage of these apps, the concept of activity time slots is introduced, based on one hour time slots. For each device and app, a time slot can either be active (if at least one request related to that app has been

[1] https://www.google.com/doubleclick/

sent during this hour) or inactive (if no activity for this app was seen during this hour). Similarly, the same concept can also be used on a category basis: a device is active for a certain category in a time slot if a request for *any* app in that category was issued during the time slot – whether multiple apps in the category generate requests is disregarded as the time slot can only be either active or inactive.

To understand the overall user behaviour in terms of app usage, the total number of users active for a particular time slot is added to get a count of application types used by the most devices over time. To further get an understanding of the diurnal usage patterns, the data can be aggregated by adding all values for the same time slot of each day.

In Figure 10, the number of users utilising at least one application of a particular category during the course of each day is shown. It is clear that regardless of the time of day, when it comes to apps that use ads, users are more likely to use their phone to play games than to do other types of activities. We cannot however be certain that this is representative of app usage in general as some popular apps are completely free without the use of ads, and other apps cost money to download and thus (in the majority of cases) do not resort to ads to generate revenue. This can for example be seen in the "Travel & Local" category, where only around $10,000$ users use ad-based apps in this category, but more than a million app downloads were seen from Google Play. In this particular case, this is due to Google Maps being one of the most downloaded apps that fall into this category and does not use ads.

VI. Discussion and Conclusions

Access to data and cloud services over cellular traffic is becoming more and more important in terms of traffic load and resource utilisation. It is also more complex than other types of access because of the heterogeneity of devices and applications. Studying and characterising the traffic allows a better understanding of the users' behaviour. This paper presents the first characterisation of such a large dataset of cellular traffic, covering millions of users.

Our study provides important pointers to network operators and cloud service providers, to optimise their network and service delivery for different types of service, such as app download [13] or video streaming [14]. Caching mechanisms can also be improved to better decide what content types to cache or even broadcast to end-user devices as proposed in [15], to optimally reduce the network load. App developers, on the other hand, may consider the daytime-dependencies for their apps' usage for proactive content pre-fetching to improve user experience. Instead of letting the developers decide whether to pre-fetch content, these mechanisms might also be triggered by the network (e.g., see [16]). Moreover, the insights into the types of devices deployed (see Section IV-C) are useful for network planning. Separating the devices according to their network capacities (in terms of down/up-link speed) allows to concentrate on higher functionality devices. Supposing that future devices have at least comparable capacities, this device category provides an outlook on future application (and network) usage.

The characterisation presented in this paper can be used for further research to explore any of these paths to inspire new algorithms and mechanisms, and evaluate them with realistic simulated traffic generated with this characterisation.

Acknowledgment

This work was partially funded by the Future Networking Solutions action line of the European Institute of Innovation and Technology ICT Labs, by the FP7 Marie Curie IRSES project MobileCloud under grant no. 612212, and was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship, co-funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant no. 246016.

References

[1] Cisco Systems, Inc., "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," Feb. 2014.

[2] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering Individual and Collective Human Dynamics from Mobile Phone Records," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, p. 224015, 2008.

[3] H. Zang and J. C. Bolot, "Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks," in *ACM Int. Conf. on Mobile Computing and Networking (MobiCom)*, 2007, pp. 123–134.

[4] Y. Zhang and A. Arvidsson, "Understanding the Characteristics of Cellular Data Traffic," in *ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design (CellNet)*, 2012.

[5] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," in *ACM Int. Conf. on Measurement and Modeling of Computer Systems (SIGMETRICS)*, 2011, pp. 305–316.

[6] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying Diverse Usage Behaviors of Smartphone Apps," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2011.

[7] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding Traffic Dynamics in Cellular Data Networks," in *IEEE INFOCOM*, 2011, pp. 882–890.

[8] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling Users in a 3G Network Using Hourglass Co-clustering," in *Int. Conf. on Mobile Computing and Networking (MobiCom)*, 2010, pp. 341–352.

[9] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2009, pp. 267–279.

[10] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in Smartphone Usage," in *Int. Conf. on Mobile Systems, Applications, and Services (MobiSys)*, 2010, pp. 179–194.

[11] Shepard, Clayton and Rahmati, Ahmad and Tossell, Chad and Zhong, Lin and Kortum, Phillip, "Livelab: Measuring wireless networks and smartphone users in the field," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 3, pp. 15–20, Jan. 2011.

[12] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer, "Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage," in *Int. Conf. on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*, 2011, pp. 47–56.

[13] F. Ben Abdesslem and A. Lindgren, "The Pursuit of 'Appiness: Exploring Android Market Download Behaviour in a Nationwide Cellular Network," in *5th International Workshop on TRaffic Analysis and Characterization (TRAC)*, 2014.

[14] ——, "Large Scale Characterisation of YouTube Requests in a Cellular Network," in *IEEE WoWMoM*, 2014.

[15] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, and Y. Grunenberger, "Is There a Case for Mobile Phone Content Pre-Staging?" in *ACM Conf. on Emerging Networking Experiments and Technologies (CoNEXT)*. ACM, 2013, pp. 321–326.

[16] B. D. Higgins, J. Flinn, T. J. Giuli, B. Noble, C. Peplin, and D. Watson, "Informed Mobile Prefetching," in *Int. Conf. on Mobile Systems, Applications, and Services (MobiSys)*, 2012, pp. 155–168.